
Discrete Restricted Boltzmann Machines

Guido F. Montúfar

Department of Mathematics
Pennsylvania State University
University Park, PA 16802
gfm10@psu.edu

Jason Morton

Department of Mathematics
Pennsylvania State University
University Park, PA 16802
morton@math.psu.edu

Abstract

We describe discrete restricted Boltzmann machines: probabilistic graphical models with bipartite interactions between discrete visible and hidden variables. These models generalize standard binary restricted Boltzmann machines and discrete naïve Bayes models. For a given number of visible variables and cardinalities of their state spaces, we bound the number of hidden variables, depending on the cardinalities of their state spaces, for which the model is a universal approximator of probability distributions. More generally, we describe exponential subfamilies and use them to bound the Kullback-Leibler approximation errors of these models from above. We use coding theory and algebraic methods to study the geometry of these models, and show that in many cases they have the dimension expected from counting parameters, but in some cases they do not. We discuss inference functions, mixtures of product distributions with shared parameters, and patterns of strong modes of probability distributions represented by discrete restricted Boltzmann machines in terms of configurations of projected products of simplices in normal fans of products of simplices.

1 Introduction

A restricted Boltzmann machine (RBM) is a probabilistic graphical model with bipartite interactions between an observed set of units and a hidden set of units (see [32, 10, 13, 14]). The RBM probability model is the set of joint probability distributions on the states of the observed units for all possible choices of interaction weights in the network. Typically RBMs are defined with binary units, but RBMs with other types of variables have also been considered, including continuous, discrete, and mixed type variables, see for instance [35, 19, 30, 8, 33]. Also probability models with more general interaction networks have been considered; including semi-restricted Boltzmann machines and higher-order interaction Boltzmann machines, see for instance [31, 20, 26, 28]. While each unit X_i of a binary RBM has the state space $\{0, 1\}$, the state space of each unit X_i of a discrete RBM is a finite set $\mathcal{X}_i = \{0, 1, \dots, r_i - 1\}$. A discrete RBM is a type of exponential family harmonium.

We discuss the representational power of discrete RBMs. We generalize previous theoretical results on standard, binary, RBMs and discrete naïve Bayes models to discrete RBMs.

A characterizing property of RBMs is that the observed units are independent given the states of the hidden units and vice versa. This is a consequence of the bipartiteness of the interaction graph, and does not depend on the units' state spaces. Discrete RBMs can be trained (in principle) using existing methods, like contrastive divergence (CD) [12, 13, 4] and expectation-maximization (EM) methods [9]. Like binary RBMs, they can be used to train the parameters of deep learning systems layer by layer [15, 3]. Compared with general network based models with hidden variables, RBMs are much more tractable, even if finding maximum likelihood estimates of target data distributions is usually difficult in either case.

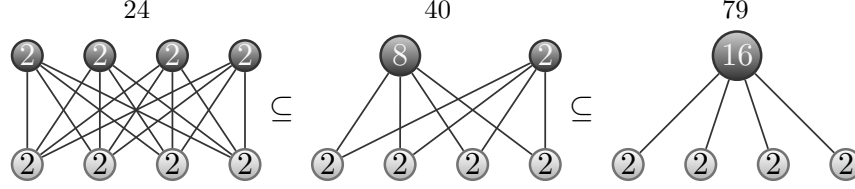


Figure 1: Examples of probability models treated in this paper, in the special case of binary visible variables. The light (dark) nodes represent visible (hidden) variables. The number inside each node indicates the cardinality of the state space of the corresponding variable. From left to right: a binary RBM; a discrete RBM with an 8-valued and a binary hidden unit; and a binary naïve Bayes model with 16 hidden classes. The total number of parameters of each model is indicated at the top.

A discrete RBM is a *product of experts* [12]. It has one expert which is a mixture model of product distributions, or naïve Bayes model, for each hidden unit. Discrete RBMs interpolate between standard binary RBMs and naïve Bayes models, which are just discrete RBMs with one single hidden unit. They can serve, in particular, to contrast distributed (restricted) mixture representations [2, 23] from binary RBMs and non-distributed (unrestricted) mixture representations from naïve Bayes models. See Figure 1.

Naïve Bayes models have been studied across many disciplines. In machine learning they are most commonly used for classification and clustering, but have also been considered for probabilistic modelling [18]. It is known that they can represent any probability distribution if the number of hidden classes is large enough, see [21] for tight bounds. In spite of their seeming simplicity, the geometry of these models is far from fully understood. Recent theoretical work on binary RBMs includes universal approximation properties [10, 16, 22], dimension and parameter identifiability [7], Bayesian learning coefficients [1], complexity [17], approximation errors [25], and distributed mixture representations [23]. We shall generalize some of these results to discrete RBMs.

Section 2 collects basic facts about independence models and hierarchical models, and briefly reviews the theory of naïve Bayes models and binary RBMs. Section 3 defines discrete RBMs formally and describes them as (i) products of mixtures of products (Proposition 8), and (ii) as restricted mixtures of products. Section 4 elaborates on the distributed mixtures of products and the inference functions represented by discrete RBMs. Proposition 12, Lemma 13, and Proposition 15 address the inference functions. Section 5 addresses the expressive power of discrete RBMs by describing tractable explicit submodels (Theorem 16) and contains results on universal approximation and maximal model approximation errors (Theorem 17). Section 6 discusses the dimension of discrete RBM models (Proposition 19 and Theorem 21). Section 7 contains an algebraic combinatorial discussion of tropicalization (Theorem 23) with consequences for the dimension of discrete RBMs collected in Propositions 26.a, 26.b, and 26.c.

2 Preliminaries

2.1 Independence models

Consider a system of $n < \infty$ random variables X_1, \dots, X_n . Assume that X_i takes states x_i in a finite set $\mathcal{X}_i = \{0, 1, \dots, r_i - 1\}$ for all $i \in \{1, \dots, n\} =: [n]$. The state space of the entire system is $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_n$. We write $x_\lambda = (x_i)_{i \in \lambda}$ for a joint state of the variables with index $i \in \lambda$ for any $\lambda \subseteq [n]$, and $x = (x_1, \dots, x_n)$ for a joint state of all variables. We denote by $\Delta(\mathcal{X})$ the set of all probability distributions on \mathcal{X} . We write $\langle a, b \rangle$ for the product $a^\top b$.

The *independence model* of X_1, \dots, X_n is the set of *product distributions* $p(x) = \prod_{i \in [n]} p_i(x_i)$ for all $x \in \mathcal{X}$, where p_i is a probability distribution on \mathcal{X}_i for all $i \in [n]$. This model is the closure in the Euclidean topology $\overline{\mathcal{E}_{\mathcal{X}}}$ of the exponential family

$$\mathcal{E}_{\mathcal{X}} = \{\exp(\langle \theta, A^{(\mathcal{X})} \rangle) : \theta \in \mathbb{R}^{d_{\mathcal{X}}}\} \quad (1)$$

with a matrix $A^{(\mathcal{X})} \in \mathbb{R}^{d_{\mathcal{X}} \times \mathcal{X}}$ of sufficient statistics $\mathbb{1}$ (constant function on \mathcal{X} with value one) and $\mathbb{1}_{\{x: x_i = y_i\}}$ for all $y_i \in \mathcal{X}_i \setminus \{0\}$ for all $i \in [n]$ (indicator functions of subsets of \mathcal{X}). The convex

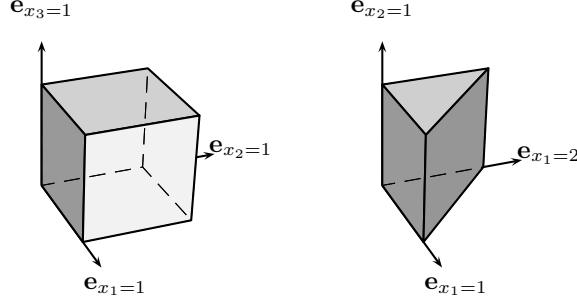


Figure 2: The convex supports of the independence models of three binary variables (left), and of two variables, one binary and one ternary (right), discussed in Example 1. Both are three-dimensional polytopes. The prism has fewer vertices than the cube and is in this sense more similar to a 3-simplex.

support of $\mathcal{E}_{\mathcal{X}}$ is the convex hull of the columns of $A^{(\mathcal{X})}$, which is a Cartesian product of simplices $Q_{\mathcal{X}} := \text{conv}(\{A_x^{(\mathcal{X})}\}_{x \in \mathcal{X}}) \cong \Delta(\mathcal{X}_1) \times \cdots \times \Delta(\mathcal{X}_n)$.

Example 1. The sufficient statistics of the independence models $\mathcal{E}_{\mathcal{X}}$ and $\mathcal{E}_{\mathcal{X}'}$ with $\mathcal{X} = \{0, 1\}^3$ and $\mathcal{X}' = \{0, 1, 2\} \times \{0, 1\}$ are, with rows labeled by indicator functions,

$$A^{(\mathcal{X})} = \left(\begin{array}{ccccccccc} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} & \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \\ \hline 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{array} \right) \begin{array}{l} x_3 = 1 \\ x_2 = 1 \\ x_1 = 2 \\ x_1 = 1 \end{array} \quad A^{(\mathcal{X}')} = \left(\begin{array}{cccccc} \begin{bmatrix} 1 \\ 2 \end{bmatrix} & \begin{bmatrix} 1 \\ 1 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 2 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ \hline 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \end{array} \right) \begin{array}{l} x_2 = 1 \\ x_1 = 2 \\ x_1 = 1 \end{array}.$$

In the first case the convex support is a cube; in the second it is a prism. See Figure 2.

2.2 Naïve Bayes models

Let $k \in \mathbb{N}$. The k -mixture of the independence model, or *naïve Bayes model* with k hidden classes, on the variables X_1, \dots, X_n is the set of all probability distributions expressible as convex combinations of k points in $\mathcal{E}_{\mathcal{X}}$:

$$\mathcal{M}_{\mathcal{X},k} := \left\{ \sum_{i \in [k]} \lambda_i p^{(i)} : \lambda_i \geq 0, p^{(i)} \in \mathcal{E}_{\mathcal{X}} \forall i \in [k], \sum_{i \in [k]} \lambda_i = 1 \right\}. \quad (2)$$

We write $\mathcal{M}_{n,k}$ for the k -mixture of the independence model of n binary variables. The dimension of the mixtures of binary product distributions are known:

Theorem 2 (Catalisano, Geramita, and Gimigliano [5]). *The mixture models of binary product distributions $\mathcal{M}_{n,k}$ have the dimension expected from counting parameters, $\min\{nk + (k-1), 2^n - 1\}$, except when $n = 4$ and $k = 3$, when $\mathcal{M}_{n,k}$ has dimension 13 instead of 14.*

Let $\mathfrak{A}(\mathcal{X}, 2)$ denote the maximal cardinality of a subset of \mathcal{X} of minimum Hamming distance at least two, i.e., the maximal cardinality of a subset $\mathcal{X}' \subseteq \mathcal{X}$ with $d_H(x, y) \geq 2$ for all distinct points $x, y \in \mathcal{X}'$, where $d_H(x, y) := |\{i \in [n] : x_i \neq y_i\}|$. The function $\mathfrak{A}_{\mathcal{X}}$ is familiar in coding theory. The k -mixtures of independence models are universal approximators when k is large enough. This can be made precise in terms of $\mathfrak{A}(\mathcal{X}, 2)$:

Theorem 3 ([21]). *The model $\overline{\mathcal{M}_{\mathcal{X},k}}$ is equal to $\Delta(\mathcal{X})$ if $k \geq \frac{|\mathcal{X}|}{\max_{i \in [n]} |\mathcal{X}_i|}$ and only if $k \geq \mathfrak{A}(\mathcal{X}, 2)$.*

When $\mathcal{X} = \{0, 1, \dots, q-1\}^n$ and q is a power of a prime number, then $\mathfrak{A}_{\mathcal{X}} = q^{n-1}$ (see [11, 34]), and by Theorem 3 $\mathcal{M}_{\mathcal{X},k} = \Delta_{\mathcal{X}}$ iff $k \geq q^{n-1}$. In particular, the smallest mixture of products model universal approximator of distributions on $\{0, 1\}^n$ has $2^{n-1}(n+1) - 1$ parameters.

A state $x \in \mathcal{X}$ is a *mode* of $p \in \Delta(\mathcal{X})$ if $p(x) > p(y)$ for all y with $d_H(x, y) = 1$. The point x is a *strong mode* of p if $p(x) > \sum_{y: d_H(x,y)=1} p(y)$.

Lemma 4 ([21]). *If a mixture $p = \sum_i \lambda_i p^{(i)}$ of product distributions has strong modes $\mathcal{C} \subseteq \mathcal{X}$, then there is a mixture component $p^{(i)}$ with mode x for each $x \in \mathcal{C}$. In particular a mixture of k product distributions has at most k strong modes.*

2.3 Binary restricted Boltzmann machines

The RBM model with n visible and m hidden binary units, denoted $\text{RBM}_{n,m}$, is the set of distributions on $\{0, 1\}^n$ of the form

$$p(x) = \frac{1}{Z} \sum_{h \in \{0,1\}^m} \exp(h^\top W x + B^\top x + C^\top h) \quad \text{for all } x \in \{0, 1\}^n, \quad (3)$$

where x denotes states of the visible units, h denotes states of the hidden units, $W = (W_{ji})_{ji} \in \mathbb{R}^{m \times n}$ is a matrix of interaction weights, $B \in \mathbb{R}^n$ and $C \in \mathbb{R}^m$ are vectors of bias weights, and $Z = \sum_{x \in \{0,1\}^n} \sum_{h \in \{0,1\}^m} \exp(h^\top W x + B^\top x + C^\top h)$ is the partition function.

It is known that binary RBMs have the expected dimension for many choices of n and m :

Theorem 5 ([7]). *The dimension of $\text{RBM}_{n,m}$ is equal to the number of parameters, $nm + n + m$, when $m + 1 \leq 2^{n - \lceil \log_2(n+1) \rceil}$, and equal to $2^n - 1$ when $m \geq 2^{n - \lceil \log_2(n+1) \rceil}$.*

It is known that binary RBMs are universal approximators when they have enough hidden units:

Theorem 6 ([22]). *The model $\overline{\text{RBM}_{n,m}}$ equals $\Delta_{\{0,1\}^n}$ whenever $m \geq 2^{n-1} - 1$.*

It is not known whether the bound from Theorem 6 is always tight, but it shows that the smallest RBM universal approximator of distributions on $\{0, 1\}^n$ has at most $2^{n-1}(n+1) - 1$ parameters, and hence not more than the smallest mixture of products universal approximator.

3 Discrete restricted Boltzmann machines

Let $\mathcal{X}_i = \{0, 1, \dots, r_i - 1\}$ for $i \in [n]$ and $\mathcal{Y}_j = \{0, 1, \dots, s_j - 1\}$ for $j \in [m]$. The graphical model with full bipartite interactions $\{\{i, j\} : i \in [n], j \in [m]\}$ on $\mathcal{X} \times \mathcal{Y}$ is the exponential family $\mathcal{E}_{\mathcal{X}, \mathcal{Y}} := \{\frac{1}{Z} \exp(\langle \theta, A^{(\mathcal{X}, \mathcal{Y})} \rangle) : \theta \in \mathbb{R}^{d_{\mathcal{X}} d_{\mathcal{Y}}}\}$ with sufficient statistics matrix equal to the Kronecker product $A^{(\mathcal{X}, \mathcal{Y})} = A^{(\mathcal{X})} \otimes A^{(\mathcal{Y})}$.

Definition 7. The *discrete RBM model* $\text{RBM}_{\mathcal{X}, \mathcal{Y}}$ is the set of marginal distributions on \mathcal{X} of $\mathcal{E}_{\mathcal{X}, \mathcal{Y}}$.

The matrix $A^{(\mathcal{X}, \mathcal{Y})}$ has $\left(\sum_{i \in [n]} (|\mathcal{X}_i| - 1) + 1\right) \left(\sum_{j \in [m]} (|\mathcal{Y}_j| - 1) + 1\right)$ linearly independent rows, and $|\mathcal{X} \times \mathcal{Y}|$ columns, corresponding to the joint states of all variables. The parametrization

$$p_\theta(x, y) = \frac{1}{Z} \exp(\langle \theta, A_{(x,y)}^{(\mathcal{X}, \mathcal{Y})} \rangle) \quad \text{for all } (x, y) \in \mathcal{X} \times \mathcal{Y} \quad (4)$$

is one-to-one, disregarding the constant row of $A^{(\mathcal{X}, \mathcal{Y})}$ which always cancels out with the normalization constant. The dimension of $\mathcal{E}_{\mathcal{X}, \mathcal{Y}}$ is equal to the number of rows of $A^{(\mathcal{X}, \mathcal{Y})}$ minus one. The dimension of $\text{RBM}_{\mathcal{X}, \mathcal{Y}}$ expected from counting parameters is equal to $\min\{\dim(\mathcal{E}_{\mathcal{X}, \mathcal{Y}}), |\mathcal{X}| - 1\}$.

In the case of one single hidden unit $\text{RBM}_{\mathcal{X}, \mathcal{Y}}$ is the naïve Bayes model on \mathcal{X} with $|\mathcal{Y}|$ hidden classes. When $\mathcal{X} = \{0, 1\}^n$ and $\mathcal{Y} = \{0, 1\}^m$, then $\text{RBM}_{\mathcal{X}, \mathcal{Y}}$ is a binary RBM. Note that $(h^\top W x + B^\top x + C^\top h) = \langle \theta, A_{(x,h)}^{(\mathcal{X}, \mathcal{Y})} \rangle$, where θ is the column by column vectorization of $\begin{pmatrix} 0 & B^\top \\ C & W \end{pmatrix}$.

Consider a parameter vector $\theta \in \mathbb{R}^{d_{\mathcal{X}} d_{\mathcal{Y}}}$ of $\mathcal{E}_{\mathcal{X}, \mathcal{Y}}$ and let $\Theta \in \mathbb{R}^{d_{\mathcal{Y}} \times d_{\mathcal{X}}}$ be a matrix with column by column vectorization equal to θ . By Roth's lemma [29] we have the identity $\theta^\top (A^{(\mathcal{X})} \otimes A^{(\mathcal{Y})})_{(x,y)} = (A_x^{(\mathcal{X})})^\top \Theta^\top A_y^{(\mathcal{Y})}$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$. This allows us to write

$$\langle \theta, A_{(x,y)}^{(\mathcal{X}, \mathcal{Y})} \rangle = \langle \Theta A_x^{(\mathcal{X})}, A_y^{(\mathcal{Y})} \rangle = \langle \Theta^\top A_y^{(\mathcal{Y})}, A_x^{(\mathcal{X})} \rangle \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}. \quad (5)$$

The inner product in eq. 5 describes following distributions:

$$p_\theta(\cdot, \cdot) = \frac{1}{Z} \exp(\langle \theta, A^{(\mathcal{X}, \mathcal{Y})} \rangle), \quad (6)$$

$$p_\theta(\cdot | x) = \frac{1}{Z_x} \exp(\langle \Theta A_x^{(\mathcal{X})}, A^{(\mathcal{Y})} \rangle), \text{ and} \quad (7)$$

$$p_\theta(\cdot | y) = \frac{1}{Z_y} \exp(\langle \Theta^\top A_y^{(\mathcal{Y})}, A^{(\mathcal{X})} \rangle). \quad (8)$$

Geometrically, $\Theta A^{(\mathcal{X})}$ is a linear projection of the columns of $A^{(\mathcal{X})}$ into the parameter space of $\mathcal{E}_\mathcal{Y}$, and similarly, $\Theta^\top A^{(\mathcal{Y})}$ is a projection into the parameter space of $\mathcal{E}_\mathcal{X}$.

Polynomial parametrization

Discrete RBMs can be parametrized not only using the exponential parametrization of hierarchical models, but also by simple polynomials.

The distributions from $\mathcal{E}_{\mathcal{X}, \mathcal{Y}}$ can be parametrized in the following way (by square free monomials):

$$p(v, h) = \frac{1}{Z} \prod_{\substack{\{j, i\} \in [m] \times [n], \\ (y'_j, x'_i) \in \mathcal{Y}_j \times \mathcal{X}_i}} (\gamma_{\{j, i\}, (y'_j, x'_i)})^{\delta_{y'_j}(h_j) \delta_{x'_i}(v_i)} \quad \forall (v, h) \in \mathcal{Y} \times \mathcal{X}, \quad (9)$$

where $\gamma_{\{j, i\}, (y'_j, x'_i)} \in \mathbb{R}_{>}$. The discrete RBM probability distributions can be written as

$$p(v) = \frac{1}{Z} \prod_{j \in [m]} \left(\sum_{h_j \in \mathcal{Y}_j} \gamma_{\{j, 1\}, (h_j, v_1)} \cdots \gamma_{\{j, n\}, (h_j, v_n)} \right) \quad \forall v \in \mathcal{X}. \quad (10)$$

Here the parameters $\gamma_{\{j, i\}, (y'_j, x'_i)}$ can be understood as $\exp(\theta_{\{j, i\}, (y'_j, x'_i)})$, where θ is a natural parameter vector of $\mathcal{E}_{\mathcal{X}, \mathcal{Y}}$.

Products of mixtures and mixtures of products

In the following we describe discrete RBMs from two complementary perspectives: (i) as products of experts, where each expert is a mixture of products, and (ii) as restricted mixtures of product distributions.

The renormalized entry-wise product (Hadamard product) of two probability distributions p and q on \mathcal{X} is defined as $p \circ q = (p(x)q(x))_{x \in \mathcal{X}} / \sum_{y \in \mathcal{X}} p(y)q(y)$.

Proposition 8. *The model $\text{RBM}_{\mathcal{X}, \mathcal{Y}}$ is a Hadamard product of mixtures of product distributions:*

$$\text{RBM}_{\mathcal{X}, \mathcal{Y}} = \mathcal{M}_{\mathcal{X}, |\mathcal{Y}_1|} \circ \cdots \circ \mathcal{M}_{\mathcal{X}, |\mathcal{Y}_m|}.$$

Proof. Proposition 8 can be seen by considering the parameterization (10). To make this explicit, one can use a *homogeneous* version of the matrix $A^{(\mathcal{X}, \mathcal{Y})}$ which we denote by A and which defines the same model. A row of A is indexed by an edge $\{i, j\}$ of the bipartite graph and a joint state $\{x_i, h_j\}$ of the visible and hidden unit connected by this edge. Such a row has a one in any column when these states agree with the global state, and zero otherwise. Let $A_{j,:}$ denote the matrix containing the rows of A with indexes $(\{i, j\}, (x_i, h_j))$ for all $x_i \in \mathcal{X}_i$ for all $i \in [n]$ for all $h_j \in \mathcal{Y}_j$, and let $A(x, h)$ denote the (x, h) -column of A . We have

$$\begin{aligned} p(x) &= \frac{1}{Z} \sum_h \exp(\langle \theta, A(x, h) \rangle) \\ &= \frac{1}{Z} \sum_h \exp(\langle \theta_{1,:}, A_{1,:}(x, h) \rangle) \exp(\langle \theta_{2,:}, A_{2,:}(x, h) \rangle) \cdots \exp(\langle \theta_{m,:}, A_{m,:}(x, h) \rangle) \\ &= \frac{1}{Z} \left(\sum_{h_1} \exp(\langle \theta_{1,:}, A_{1,:}(x, h_1) \rangle) \right) \cdots \left(\sum_{h_m} \exp(\langle \theta_{m,:}, A_{m,:}(x, h_m) \rangle) \right) \\ &= \frac{1}{Z} (Z_1 p^{(1)}(x)) \cdots (Z_m p^{(m)}(x)) = \frac{1}{Z} p^{(1)}(x) \cdots p^{(m)}(x), \end{aligned}$$

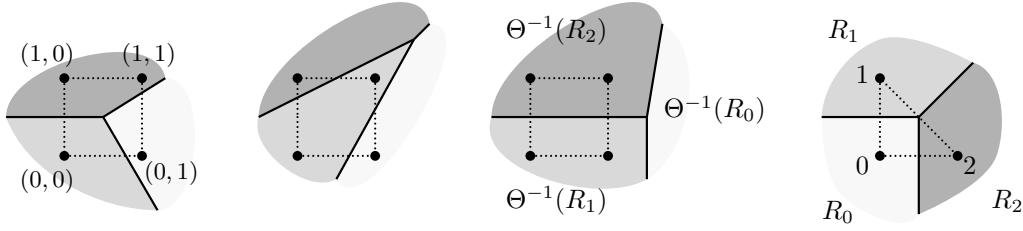


Figure 3: Three 3-slicings of the 2-cube by the fan of the 2-simplex with maximal cones R_0 , R_1 , and R_2 . Each vertex of the 2-cube is a column vector of the sufficient statistics matrix of the 2-bit independence model. Each vertex of the 2-simplex is a column vector of the sufficient statistics matrix of the independence model of one single ternary variable (equal to $\Delta(\{0, 1, 2\})$).

where $p^{(j)} \in \mathcal{M}_{\mathcal{X}, |\mathcal{Y}_j|}$ and $Z_j = \sum_{x \in \mathcal{X}} \sum_{h_j \in \mathcal{Y}_j} \exp(\langle \theta_{j,:}, A_{j,:}(x, h_j) \rangle)$ for all $j \in [m]$. Since the vectors $\theta_{j,:}$ can be chosen arbitrarily, the factors $p^{(j)}$ can be made arbitrary within $\mathcal{M}_{\mathcal{X}, |\mathcal{Y}_j|}$. \square

Of course, every distribution in $\text{RBM}_{\mathcal{X}, \mathcal{Y}}$ is a mixture distribution $p(x) = \sum_{h \in \mathcal{Y}} p(x|h)q(h)$. The mixture weights are given by the marginals $q(h)$ on \mathcal{Y} of distributions from $\mathcal{E}_{\mathcal{X}, \mathcal{Y}}$, and the mixture components can be described as follows:

Proposition 9. *The set of conditional distributions $p(x|h)$, $h \in \mathcal{Y}$ of $\mathcal{E}_{\mathcal{X}, \mathcal{Y}}$ is the set of product distributions in $\mathcal{E}_{\mathcal{X}}$ with parameters $\theta_h = \Theta^\top A_h^{(\mathcal{Y})}$, $h \in \mathcal{Y}$ equal to a linear projection of the vertices $\{A_h^{(\mathcal{Y})} : h \in \mathcal{Y}\}$ of the Cartesian product of simplices $Q_{\mathcal{Y}} \cong \Delta(\mathcal{Y}_1) \times \cdots \times \Delta(\mathcal{Y}_m)$.*

Proof. This is by eq. 5. \square

4 Products of simplices and their normal fans

Binary RBMs have been analyzed by considering each of the m hidden units as defining a hyperplane H_j slicing the n -cube into two regions. To generalize the results provided by this analysis, in this section we replace the n -cube with a general product of simplices $Q_{\mathcal{X}}$, and replace the two regions defined by the hyperplane H_j by the $|\mathcal{Y}_j|$ regions defined by maximal cones of the normal fan of the simplex $\Delta(\mathcal{Y}_j)$.

Subdivisions of independence models

The *normal cone* of a polytope $Q \subset \mathbb{R}^d$ at a point $x \in Q$ is the set of all vectors $v \in \mathbb{R}^d$ with $\langle v, (x - y) \rangle \geq 0$ for all $y \in Q$. We denote by R_x the normal cone of the product of simplices $Q_{\mathcal{X}} = \text{conv}\{A_x^{(\mathcal{X})}\}_{x \in \mathcal{X}}$ at the vertex $A_x^{(\mathcal{X})}$. The *normal fan* $\mathcal{F}_{\mathcal{X}}$ is the collection of all normal cones of $Q_{\mathcal{X}}$.

The product distributions $p_\theta = \frac{1}{Z} \exp(\langle \theta, A^{(\mathcal{X})} \rangle) \in \mathcal{E}_{\mathcal{X}}$ strictly maximized at $x \in \mathcal{X}$, i.e., which satisfy $p_\theta(x) > p_\theta(y) \forall y \in \mathcal{X} \setminus \{x\}$, are those with parameters θ in the relative interior of R_x .

Inference functions and slicings

For any choice of parameters of the model $\text{RBM}_{\mathcal{X}, \mathcal{Y}}$, there is an *inference function* $\pi: \mathcal{X} \rightarrow \mathcal{Y}$, (or more generally $\pi: \mathcal{X} \rightarrow 2^{\mathcal{Y}}$), which computes the most likely hidden state given a visible state. These functions are not necessarily injective nor surjective. For a visible state x , the conditional distribution on the hidden states is a product distribution $p(y|X = x) = \frac{1}{Z} \exp(\langle \Theta A_x^{(\mathcal{X})}, A_y^{(\mathcal{Y})} \rangle)$, which is maximized at the y for which $\Theta A_x^{(\mathcal{X})} \in R_y$. The preimages of the R_y by Θ partition the input space $\mathbb{R}^{d_{\mathcal{X}}}$, and are called *inference regions*. See Figure 3 and Example 11.

Definition 10. A \mathcal{Y} -slicing of a finite set \mathcal{Z} is a partition of \mathcal{Z} into the preimages of R_y , the maximal cones of $\mathcal{F}_{\mathcal{Y}}$, by a linear map Θ . We assume that Θ is generic, such that it maps each point in \mathcal{Z} into the interior of some R_y .

When $\mathcal{Y} = \{0, 1\}$, the fan $\mathcal{F}_{\mathcal{Y}}$ consists of a hyperplane and the two closed halfspaces defined by that hyperplane. A \mathcal{Y} -slicing is in this case a standard slicing by a hyperplane.

Example 11. Let $\mathcal{X} = \{0, 1, 2\} \times \{0, 1\}$ and $\mathcal{Y} = \{0, 1\}^4$. The maximal cones R_y of the normal fan of the 4-cube with vertices $\{0, 1\}^4$ are the closed orthants of \mathbb{R}^4 . The 6 vertices $\{A_x^{(\mathcal{X})} : x \in \mathcal{X}\}$ of the prism $\Delta_2 \times \Delta_1$ can be mapped into 6 distinct orthants of \mathbb{R}^4 each with an even number of positive coordinates:

$$\begin{pmatrix} 3 & -2 & -2 & -2 \\ 1 & 2 & -2 & -2 \\ 1 & -2 & -2 & 2 \\ 1 & -2 & 2 & -2 \end{pmatrix}_{\Theta} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}_{A^{(\mathcal{X})}} = \begin{pmatrix} -1 & -1 & 1 & 1 & 1 & 3 \\ 1 & 1 & 3 & -1 & -1 & 1 \\ -3 & 1 & -1 & -1 & 3 & 1 \\ 1 & -3 & -1 & 3 & -1 & 1 \end{pmatrix}. \quad (11)$$

Even in the case of one hidden unit, the slicings can be complex, but the following simple type of slicing is always available.

Proposition 12. Any slicing by $k - 1$ parallel hyperplanes is a $\{1, 2, \dots, k\}$ -slicing.

Proof. We show that there is a line $\mathcal{L} = \{\lambda r - b : \lambda \in \mathbb{R}\}$, $r, b \in \mathbb{R}^k$ intersecting all cells of $\mathcal{F}_{\mathcal{Y}}$, $\mathcal{Y} = \{1, \dots, k\}$. We need to show that there is a choice of r and b such that for every $y \in \mathcal{Y}$ the set $I_y \subseteq \mathbb{R}$ of all λ with $\langle \lambda r - b, (e_y - e_z) \rangle > 0$ for all $z \in \mathcal{Y} \setminus \{y\}$ has a non-empty interior. Now, I_y is the set of λ with

$$\lambda(r_y - r_z) > b_y - b_z \quad \text{for all } z \neq y. \quad (12)$$

Choosing $b_1 < \dots < b_k$ and $r_y = f(b_y)$, where f is a strictly increasing and strictly concave function, we get $I_1 = (-\infty, \frac{b_2 - b_1}{r_2 - r_1})$, $I_y = (\frac{b_y - b_{y-1}}{r_y - r_{y-1}}, \frac{b_{y+1} - b_y}{r_{y+1} - r_y})$ for $y = 2, 3, \dots, k - 1$, and $I_k = (\frac{b_k - b_{k-1}}{r_k - r_{k-1}}, \infty)$. The lengths $\infty, l_2, \dots, l_{k-1}, \infty$ of the intervals I_1, \dots, I_k can be adjusted arbitrarily by choosing suitable differences $r_{j+1} - r_j$ for all $j = 1, \dots, k - 1$. \square

Strong modes

A strong mode of a distribution p on \mathcal{X} is a point $x \in \mathcal{X}$ such that $p(x) > \sum_{y \in d_H(x, y) = 1} p(y)$, where $d_H(x, y)$ is the Hamming distance between x and y .

Lemma 13. Let $\mathcal{C} \subseteq \mathcal{X}$ be a set of arrays which are pairwise different in at least two entries (a code of minimum distance two). If $\text{RBM}_{\mathcal{X}, \mathcal{Y}}$ contains a probability distribution with strong modes \mathcal{C} , then there is a linear map of $\{A_y^{(\mathcal{Y})}\}$ into the \mathcal{C} -cells (the cones above the codewords in the normal fan) of $\mathcal{F}_{\mathcal{X}}$ sending at least one vertex into each cell.

On the other hand, let $\mathcal{C} \subseteq \mathcal{X}$. If there is a linear map Θ of the vertices of $\times_{j \in [m]} \Delta_{\mathcal{Y}_j}$ into the \mathcal{C} -cells R_x , $x \in \mathcal{C}$ of the fan $\mathcal{F}_{\mathcal{X}}$, with $\max_x \{\langle \Theta^\top A_y^{(\mathcal{Y})}, A_x^{(\mathcal{X})} \rangle\} = c$ for all y , then $\text{RBM}_{\mathcal{X}, \mathcal{Y}}$ contains a probability distribution with strong modes \mathcal{C} .

Proof. This is by Proposition 9, Lemma 4 and the definition of the normal fan. \square

By this lemma, if $\text{RBM}_{\mathcal{X}, \mathcal{Y}}$ is a universal approximator of distributions on \mathcal{X} , then $|\mathcal{Y}| \geq \mathfrak{A}(\mathcal{X}, 2)$. Hence discrete RBMs may not be universal approximators even when they have the same dimension as the ambient probability simplex.

Example 14. Let $\mathcal{X} = \{0, 1, 2\}^n$ and $\mathcal{Y} = \{0, 1, \dots, 4\}^m$. In this case $\mathfrak{A}(\mathcal{X}, 2) = 3^{n-1}$. When $n = 3$ or $n = 4$, if $\text{RBM}_{\mathcal{X}, \mathcal{Y}}$ is a universal approximator, then $m \geq 2$ and $m \geq 3$. On the other hand, the smallest m for which $\text{RBM}_{\mathcal{X}, \mathcal{Y}}$ has $3^n - 1$ parameters is $m = 1$ and $m = 2$.

Using the analysis of [23] gives the following.

Proposition 15. If $4\lceil m/3 \rceil \leq n$, then $\text{RBM}_{\mathcal{X}, \mathcal{Y}}$ contains distributions with 2^m strong modes.

5 Approximation errors and universal approximation

In this section we describe certain explicit tractable submodels of the discrete RBM and use these to provide error bounds.

Theorem 16. *Let $d_Y = 1 + \sum_{j=1}^m (|\mathcal{Y}_j| - 1)$. The model $\text{RBM}_{\mathcal{X}, \mathcal{Y}}$ can approximate the following arbitrarily well:*

- Any mixture of d_Y product distributions with disjoint supports.
- When $d_Y \geq (\prod_{i \in [k]} |\mathcal{X}_i|) / \max_{j \in [k]} |\mathcal{X}_j|$, any distribution from the set \mathcal{P} of distributions with constant value on each block $\{x_1\} \times \dots \times \{x_k\} \times \mathcal{X}_{k+1} \times \dots \times \mathcal{X}_n$ for all $x_i \in \mathcal{X}_i$, for all $i \in [k]$.
- Any probability distribution with support contained in the union of d_Y sets of the form $\{x_1\} \times \dots \times \{x_{k-1}\} \times \mathcal{X}_k \times \{x_{k+1}\} \times \dots \times \{x_n\}$.

Proof. By Proposition 8 the RBM model contains any product $p^{(1)} \circ \dots \circ p^{(m)}$, where $p^{(j)} \in \mathcal{M}_{\mathcal{X}, |\mathcal{Y}_j|}$ for all $j \in [m]$. In particular, it contains $p = p^{(0)} \circ (\mathbb{1} + \tilde{\lambda}_1 \tilde{p}^{(1)}) \circ \dots \circ (\mathbb{1} + \tilde{\lambda}_m \tilde{p}^{(m)})$, where $p^{(0)} \in \mathcal{E}_{\mathcal{X}}$ and $\tilde{p}^{(j)} \in \mathcal{M}_{\mathcal{X}, |\mathcal{Y}_j| - 1}$. Choosing the factors $\tilde{p}^{(j)}$ with disjoint supports results in $p = \sum_{j=0}^m \lambda_j p^{(j)}$, where $p^{(0)}$ is any product distribution and $p^{(j)} \in \mathcal{M}_{\mathcal{X}, |\mathcal{Y}_j| - 1}$ can be made arbitrary for all $j \in [m]$, as long as $\text{supp}(p^{(j)}) \cap \text{supp}(p^{(j')}) = \emptyset$ for all $j \neq j'$.

The second item: Any point in \mathcal{P} is a mixture of the uniform distributions p_{x_1, \dots, x_k} on the blocks $\{x_1\} \times \dots \times \{x_k\} \times \mathcal{X}_{k+1} \times \dots \times \mathcal{X}_n$. These mixture components have disjoint supports and are product distributions, since they factorize as $p_{x_1, \dots, x_k} = \prod_{i \in [k]} \delta_{x_i} \prod_{i \in [n] \setminus [k]} u_i$, where u_i denotes the uniform distribution on \mathcal{X}_i . For any $j \in [k]$, any mixture of the form $\sum_{x_j \in \mathcal{X}_j} \lambda_{x_j} p_{x_1, \dots, x_k}$ is also a product distribution which factorizes as

$$\left(\sum_{x_j \in \mathcal{X}_j} \lambda_{x_j} \delta_{x_j} \right) \prod_{i \in [k] \setminus \{j\}} \delta_{x_i} \prod_{i \in [n] \setminus [k]} u_i. \quad (13)$$

Hence any point in \mathcal{P} is a mixture of $(\prod_{i \in [k]} |\mathcal{X}_i|) / \max_{j \in [k]} |\mathcal{X}_j|$ product distributions with disjoint supports.

The third item follows from the first item, because $\overline{\mathcal{E}_{\mathcal{X}}}$ contains any distribution with support of the form $\{x_1\} \times \dots \times \{x_{k-1}\} \times \mathcal{X}_k \times \{x_{k+1}\} \times \dots \times \{x_n\}$. See [21]. \square

Let $p, q \in \Delta(\mathcal{X})$. The Kullback-Leibler (KL) divergence from q to p is defined as $D(q||p) := \sum_{x \in \mathcal{X}} q(x) \log \frac{q(x)}{p(x)}$ when $\text{supp}(p) \supseteq \text{supp}(q)$ and as $D(p||q) = \infty$ otherwise. The divergence from p to a model $\mathcal{M} \subseteq \Delta(\mathcal{X})$ is defined as $D(p||\mathcal{M}) := \inf_{q \in \mathcal{M}} D(p||q)$. A model \mathcal{M} of distributions on \mathcal{X} is a *universal approximator* iff $D(p||\mathcal{M}) = 0$ for all $p \in \Delta(\mathcal{X})$.

Theorem 17. *Let $\Lambda \subseteq [n]$. If $\prod_{i \in [n] \setminus \Lambda} |\mathcal{X}_i| \leq 1 + \sum_{j \in [m]} (|\mathcal{Y}_j| - 1) = d_Y$, then the KL-divergence from any $p \in \Delta(\mathcal{X})$ to the model $\text{RBM}_{\mathcal{X}, \mathcal{Y}}$ is bounded by*

$$D(p||\text{RBM}_{\mathcal{X}, \mathcal{Y}}) \leq \log \frac{\prod_{i \in \Lambda} |\mathcal{X}_i|}{\max_{i \in \Lambda} |\mathcal{X}_i|}.$$

In particular, $\text{RBM}_{\mathcal{X}, \mathcal{Y}}$ is a universal approximator whenever $d_Y \geq |\mathcal{X}| / \max_{i \in [n]} |\mathcal{X}_i|$.

Proof. The set \mathcal{P} described in the second item of Theorem 16 is known as a *partition model*. The maximal divergence from such a model is equal to the logarithm of the cardinality of the largest block. See [25]. We have thus $\max_p D(p||\text{RBM}_{\mathcal{X}, \mathcal{Y}}) \leq \max_p D(p||\mathcal{P}) = \log \frac{\prod_{i \in \Lambda} |\mathcal{X}_i|}{\max_{i \in \Lambda} |\mathcal{X}_i|}$. \square

This theorem tells us that the maximal approximation error increases at most logarithmically with the total number of visible states, and decreases at least logarithmically with the sum of the number of states of the hidden units. This observation could be helpful, for example, in designing a penalty term to allow comparison of models with differing numbers of units.

Remark 18. The submodels of discrete RBMs described in Lemma 16 can also be used to upper bound the expected Kullback-Leibler divergence from q to $\text{RBM}_{\mathcal{X},\mathcal{Y}}$ when q is drawn from a prior on the probability simplex $\Delta(\mathcal{X})$. The expectation value of the divergence for such submodels and any Dirichlet priors has been computed in [24].

6 Dimension

In this section we study the dimension of the models $\text{RBM}_{\mathcal{X},\mathcal{Y}}$. Our analysis builds on previous work by Cueto, Morton, and Sturmfels [7], where the binary case was treated. The idea is to bound the dimension from below by the dimension of a related max-plus model, called the tropical RBM [27], and from above by the dimension expected from counting parameters. One reason RBMs are attractive is that they have a large learning capacity, e.g. may be built with millions of parameters. Dimension calculations show whether those parameters are wasted, or translate into higher-dimensional spaces of representable distributions.

The dimension of the discrete RBM model can be bounded from above not only by its expected dimension, but also by a function of the dimension of its Hadamard factors:

Proposition 19. *The dimension of the discrete RBM is bounded as*

$$\dim(\text{RBM}_{\mathcal{X},\mathcal{Y}}) \leq \dim(\mathcal{M}_{\mathcal{X},|\mathcal{Y}_i|}) + \sum_{j \in [m] \setminus \{i\}} \dim(\mathcal{M}_{\mathcal{X},|\mathcal{Y}_j|-1}) + (m-1) \quad \text{for all } i \in [m].$$

Hence $\text{RBM}_{\mathcal{X},\mathcal{Y}}$ can have the expected dimension only if (i) the right hand side equals $|\mathcal{X}| - 1$, or (ii) each mixture model $\mathcal{M}_{\mathcal{X},k}$ has the expected dimension for all $k = 1, \dots, \max_j |\mathcal{Y}_j|$.

Proof. Note that $\mathcal{E}_{\mathcal{X}} \circ \mathcal{E}_{\mathcal{X}} = \mathcal{E}_{\mathcal{X}}$ and hence $\mathcal{E}_{\mathcal{X}} \circ \mathcal{M}_{\mathcal{X},k} = \mathcal{M}_{\mathcal{X},k}$. Let u denote the uniform distribution. By Proposition 8

$$\text{RBM}_{\mathcal{X},\mathcal{Y}} = (\mathcal{M}_{\mathcal{X},|\mathcal{Y}_1|}) \circ (\lambda_1 u + (1 - \lambda_1) \mathcal{M}_{\mathcal{X},|\mathcal{Y}_1|}) \circ \dots \circ (\lambda_m u + (1 - \lambda_m) \mathcal{M}_{\mathcal{X},|\mathcal{Y}_m|-1}),$$

from which the claim follows. \square

Example 20. Consider an RBM with only two visible variables. The set of $M \times N$ matrices of rank at most k has dimension $k(M + N - k)$ for all $1 \leq k < \min\{M, N\}$. Hence the k -mixture of the independence model of two variables has dimension less than the number of parameters whenever $1 < k < \min\{|\mathcal{X}_1|, |\mathcal{X}_2|\}$.

By Proposition 19, if $(\sum_{j \in [m]} (|\mathcal{Y}_j| - 1) + 1)(|\mathcal{X}_1| + |\mathcal{X}_2| - 1) \leq |\mathcal{X}_1 \times \mathcal{X}_2|$ and $1 < |\mathcal{Y}_j| \leq \min\{|\mathcal{X}_1|, |\mathcal{X}_2|\}$ for some $j \in [m]$, then $\text{RBM}_{\mathcal{X}_1 \times \mathcal{X}_2, \mathcal{Y}}$ does *not* have the expected dimension.

We say that a set $\mathcal{Z} \subset \mathcal{X}$ is full rank when the matrix with columns $\{A_x^{(\mathcal{X})} : x \in \mathcal{Z}\}$ has full rank.

Theorem 21. *The model $\text{RBM}_{\mathcal{X},\mathcal{Y}}$ has dimension $(1 + \sum_{i \in [n]} (|\mathcal{X}_i| - 1))(1 + \sum_{j \in [m]} (|\mathcal{Y}_j| - 1)) - 1$, as expected from counting parameters, whenever \mathcal{X} contains m disjoint Hamming balls of radii $2(|\mathcal{Y}_j| - 1) - 1$, $j \in [m]$ and the subset of \mathcal{X} not contained in these balls has full rank. On the other hand, if m Hamming balls of radius one cover \mathcal{X} , then $\dim(\text{RBM}_{\mathcal{X},\mathcal{Y}}) = |\mathcal{X}| - 1$.*

In order to prove this theorem we will need two main tools: slicings by normal fans of simplices (Section 4), and the tropical RBM model, described in Section 7. The theorem will follow from the analysis contained in the next section.

7 Tropical model

Let $p_{\theta}(v) = \sum_h p_{\theta}(v, h) = \frac{1}{Z} \sum_h \exp(\langle \theta, A_{(v,h)}^{(\mathcal{X},\mathcal{Y})} \rangle)$, $\theta \in \mathbb{R}^d$ be a parametrization of $\text{RBM}_{\mathcal{X},\mathcal{Y}}$.

Definition 22. The tropical model $\text{RBM}_{\mathcal{X},\mathcal{Y}}^{\text{tropical}}$ is the image of the tropical morphism Φ , which evaluates $\log(p_{\theta}(v)) = \log(\sum_h p_{\theta}(v, h))$ for all $v \in \mathcal{X}$ and $\theta \in \mathbb{R}^d$ within the max-plus algebra (addition becomes $a + b = \max\{a, b\}$) and only up to additive constants independent of v (i.e., disregarding the normalization constant Z).

The idea behind this definition is that $\log(\exp(a) + \exp(b)) \approx \max\{a, b\}$, when a and b have a different order of magnitude. We have

$$\Phi(v; \theta) = \max\{\langle \theta, A_{(v,h)}^{(\mathcal{X}, \mathcal{Y})} \rangle : h \in \mathcal{Y}\} \quad \text{for all } v \in \mathcal{X}, \theta \in \mathbb{R}^d. \quad (14)$$

The tropical model captures important properties of the original model. Of particular interest is the relation

$$\dim(\text{RBM}_{\mathcal{X}, \mathcal{Y}}^{\text{tropical}}) \leq \dim(\text{RBM}_{\mathcal{X}, \mathcal{Y}}) \leq \min\{\dim(\mathcal{E}_{\mathcal{X}, \mathcal{Y}}), |\mathcal{X}| - 1\}, \quad (15)$$

which gives us a tool to estimate the dimension of the discrete RBM model.

The following Theorem 23 describes the regions of linearity of Φ . This allows us to express the dimension of $\text{RBM}_{\mathcal{X}, \mathcal{Y}}^{\text{tropical}}$ as the maximum rank of a class of matrices defined by \mathcal{Y}_j -slicings (see Definition 10) of the set $\{A_v^{(\mathcal{X})}\}_v$ for all $j \in [m]$.

For each $j \in [m]$, let $C_j = \{C_{j,1}, \dots, C_{j,|\mathcal{Y}_j|}\}$ be a \mathcal{Y}_j -slicing of $\{A_x^{(\mathcal{X})} : x \in \mathcal{X}\}$. Let $A_{C_{j,k}}^\top$ be of $A^\mathcal{X}$ with the columns corresponding to points not in $C_{j,k}$ zeroed and $A_{C_j} = (A_{C_{j,1}} | \dots | A_{C_{j,|\mathcal{Y}_j|}})$. The matrix A_{C_j} is $|\mathcal{X}| \times |\mathcal{Y}_j|d\mathcal{X}$. Let $d = \sum_{j \in [m]} |\mathcal{Y}_j|d\mathcal{X}$.

Theorem 23. *On each region of linearity corresponding to a collection of m \mathcal{Y}_j -slicings, the tropical morphism $\Phi: \mathbb{R}^d \rightarrow \text{RBM}_{\mathcal{X}, \mathcal{Y}}^{\text{tropical}}$ is the linear map represented by the $|\mathcal{X}| \times d$ -matrix*

$$\mathcal{A} = (A_{C_1} | \dots | A_{C_m}),$$

modulo constant functions. In particular $\dim(\text{RBM}_{\mathcal{X}, \mathcal{Y}}^{\text{tropical}}) + 1$ is the maximum rank of \mathcal{A} over all possible collections of m \mathcal{Y}_j -slicings.

Proof. Again use the homogeneous version of the matrix $A^{(\mathcal{X}, \mathcal{Y})}$ as in the proof of Proposition 8; this will not affect the rank of \mathcal{A} . Let $\theta_{h_j} = (\theta_{\{j,i\}, (h_j, x_i)})_{i \in [n], x_i \in \mathcal{X}_i}$ and denote by A_{h_j} the submatrix of $A^{(\mathcal{X}, \mathcal{Y})}$ containing the rows with indices $\{\{j,i\}, (h_j, x_i) : i \in [n], x_i \in \mathcal{X}_i\}$. For a given $v \in \mathcal{X}$ we have

$$\max\{\langle \theta, A_{(v,h)}^{(\mathcal{X}, \mathcal{Y})} \rangle : h \in \mathcal{Y}\} = \sum_{j \in [m]} \max\{\langle \theta_{h_j}, A_{h_j}(v, h_j) \rangle : h_j \in \mathcal{Y}_j\}. \quad \square$$

In the following we evaluate the maximum rank of the matrix \mathcal{A} for various choices of \mathcal{X} and \mathcal{Y} by examining good slicings.

Lemma 24. *For any $x^* \in \mathcal{X}$ and $0 < k < n$ the affine hull of the set $\{A_x^{(\mathcal{X})} : d_H(x, x^*) = k\}$ has dimension $\sum_{i \in [n]} (|\mathcal{X}_i| - 1) - 1$.*

Proof. The set $\mathcal{Z}^k := \{A_x^{(\mathcal{X})} : d_H(x, x^*) = k\}$ is the subset of vertices of the product of simplices $Q_{\mathcal{X}}$ contained in the hyperplane $H^k := \{z : \langle \mathbf{1}, z \rangle = k + 1\}$. We have that $\text{conv}(\mathcal{Z}^k) = Q_{\mathcal{X}} \cap H^k$, because if not, H^k would slice an edge of $Q_{\mathcal{X}}$. On the other hand the two vertices of any edge of $Q_{\mathcal{X}}$ lie in two parallel hyperplanes H^l and H^{l+1} , and hence $\langle \mathbf{1}, z \rangle \notin \mathbb{N}$ for any point z in the relative interior of an edge of $Q_{\mathcal{X}}$. The set \mathcal{Z}^k is not contained in any proper face of $Q_{\mathcal{X}}$ and hence $\text{conv}(\mathcal{Z}^k)$ intersects the interior of $Q_{\mathcal{X}}$. Thus $\dim(\text{conv}(\mathcal{Z}^k)) = \dim(Q_{\mathcal{X}}) - 1$, as was claimed. \square

If \mathcal{Z} is a radius-one Hamming ball in \mathcal{X} , then the $1 + \sum_{i \in [n]} (|\mathcal{X}_i| - 1)$ vectors $A_x^{(\mathcal{X})}$, $x \in \mathcal{Z}$ are affinely independent. Lemma 24 implies the following.

Corollary 25. *Let $x \in \mathcal{X}$, and $2k - 3 \leq n$. There is a slicing $C_1 = \{C_{1,1}, \dots, C_{1,k}\}$ of \mathcal{X} by $k - 1$ parallel hyperplanes such that $\bigcup_{l=1}^{k-1} C_{1,l} = B_x(2k - 3)$ is the Hamming ball of radius $2k - 3$ centered at x , and $A_{C_1} = (A_{C_{1,1}} | \dots | A_{C_{1,k-1}})$ has full rank.*

Recall that $\mathfrak{A}(\mathcal{X}, d)$ denotes the maximal cardinality of a subset of \mathcal{X} of minimum Hamming distance at least d . When $\mathcal{X} = \{0, 1, \dots, q - 1\}$ we write $\mathfrak{A}_q(n, d)$. Let $\mathfrak{R}(\mathcal{X}, d)$ denote the minimal cardinality of a subset of \mathcal{X} with covering radius d .

Proposition 26.a (Binary visible units). *Let $\mathcal{X} = \{0, 1\}^n$ and $|\mathcal{Y}_j| = s_j$, $j \in [m]$. If \mathcal{X} contains m disjoint Hamming balls of radii $2s_j - 3$, $j \in [m]$ whose complement has maximum rank, then $\text{RBM}_{\mathcal{X}, \mathcal{Y}}^{\text{tropical}}$ has the expected dimension, $\min\{\sum_{j \in [m]} (s_j - 1)(n + 1) + n, 2^n - 1\}$.*

In particular, if $\mathcal{Y} = \{0, 1, \dots, s - 1\}^m$ and $m < \mathfrak{A}_2(n, d)$, $d = 4(s - 1) - 1$, then $\text{RBM}_{\mathcal{X}, \mathcal{Y}}$ has the expected dimension. It is known that $\mathfrak{A}_2(n, d) \geq 2^{n - \lceil \log_2(\sum_{j=0}^{d-2} \binom{n-1}{j}) \rceil}$.

Proposition 26.b (Binary hidden units). *Let $\mathcal{Y} = \{0, 1\}^m$.*

- *If $m + 1 \leq \mathfrak{A}(\mathcal{X}, 3)$, then $\text{RBM}_{\mathcal{X}, \{0, 1\}^m}^{\text{tropical}}$ has dimension $(1 + m)(1 + \sum_{i \in [n]} (|\mathcal{X}_i| - 1)) - 1$.*
- *If $m + 1 \geq \mathfrak{K}(\mathcal{X}, 1)$, then $\text{RBM}_{\mathcal{X}, \{0, 1\}^m}^{\text{tropical}}$ has dimension $|\mathcal{X}| - 1$.*

Let $\mathcal{Y} = \{0, 1\}^m$ and $\mathcal{X} = \{0, 1, \dots, q - 1\}^n$, where q is a prime power.

- *If $m + 1 \leq q^{n - \lceil \log_q(1 + (n-1)(q-1) + 1) \rceil}$, then $\text{RBM}_{\mathcal{X}, \mathcal{Y}}^{\text{tropical}}$ has dimension $(1 + m)(1 + \sum_{i \in [n]} (|\mathcal{X}_i| - 1)) - 1$.*
- *If $n = (q^r - 1)/(q - 1)$ for some $r \geq 2$, then $\mathcal{A}_{\mathcal{X}}(3) = \mathfrak{K}(\mathcal{X}, 1)$, and $\text{RBM}_{\mathcal{X}, \mathcal{Y}}^{\text{tropical}}$ has the expected dimension for any m .*

When both hidden and visible units are binary and $m < 2^{n - \lceil \log_2(n+1) \rceil}$, then $\text{RBM}_{\mathcal{X}, \mathcal{Y}}$ has the expected dimension.

Proposition 26.c (Arbitrary sized units). *If \mathcal{X} contains m disjoint Hamming balls of radii $2|\mathcal{Y}_1| - 3, \dots, 2|\mathcal{Y}_m| - 3$, and the complement of their union has full rank, then $\text{RBM}_{\mathcal{X}, \mathcal{Y}}^{\text{tropical}}$ has the expected dimension.*

Proof. Propositions 26.a, 26.b, and 26.c follow from Theorem 23 and Corollary 25 together with the following explicit bounds on \mathfrak{A} .

The q -ary Hamming codes are perfect linear codes over the finite field \mathbb{F}_q of length $n = (q^r - 1)/(q - 1)$ minimum Hamming distance three and covering radius one.

$\mathfrak{A}_q(n, d) \geq \frac{q^n}{\sum_{j=0}^{d-1} \binom{n}{j} (q-1)^j}$. Furthermore, if q is a prime power, $\mathfrak{A}_q(n, d) \geq q^k$, where k is the largest integer with $q^k < \frac{q^n}{\sum_{j=0}^{d-2} \binom{n-1}{j} (q-1)^j}$ (Gilbert-Varshamov [11, 34]). In particular, $\mathfrak{A}_2(n, 3) \geq 2^k$, where k is the largest integer with $2^k < \frac{2^n}{(n-1)+1} = 2^{n - \log_2(n)}$, i.e., $k = n - \lceil \log_2(n+1) \rceil$. \square

Example 27. Many results in coding theory can now be translated directly to a statement about the dimension of discrete RBMs. Here is an example. Let $\mathcal{X} = \{1, 2, \dots, s\} \times \{1, 2, \dots, s\} \times \{1, 2, \dots, t\}$, $s \leq t$. The minimum cardinality $\mathfrak{K}(\mathcal{X}, 1)$ of a code $C \subseteq \mathcal{X}$ with covering radius one equals $s^2 - \lfloor \frac{(3s-t)^2}{8} \rfloor$ if $t \leq 3s$, and s^2 otherwise, see [6, Theorem 3.7.4]. Hence $\text{RBM}_{\mathcal{X}, \{0, 1\}^m}$ has dimension $|\mathcal{X}| - 1$ when $m + 1 \geq s^2 - \lfloor \frac{(3s-t)^2}{8} \rfloor$ and $t \leq 3s$, and when $m + 1 \geq s^2$ and $t > 3s$.

8 Discussion

We generalize various theoretical results on binary RBMs and naïve Bayes models to the more general class of discrete RBMs. We highlight and contrast geometric and combinatorial properties of distributed products of experts and non-distributed mixtures of experts. In particular, we estimate the number of hidden units for which these models are universal approximators, depending on the cardinalities of the state spaces of all units. Moreover, we show that the maximal Kullback-Leibler approximation errors of these models are bounded from above by the expression $\log(\prod_{i \in [n]} |\mathcal{X}_i|) - \log(\max_i |\mathcal{X}_i|) - \log(\sum_{j \in [m]} (|\mathcal{Y}_j| - 1))$, respectively vanish when the expression is not positive (Theorem 17). This generalizes [25, Theorem 5.1], which states that the maximal divergence to the binary model decreases at least logarithmically with the number of hidden units. And it shows that the maximal approximation error decreases at least logarithmically in the number of states of each

hidden node. We computed exponential subfamilies of $\text{RBM}_{\mathcal{X},\mathcal{Y}}$ which can be used to estimate the expected approximation errors as well. We discussed inference functions of these models in terms of normal fans of products of simplices.

We discuss the combinatorics of the tropical versions of discrete RBMs, and use this to show that the model $\text{RBM}_{\mathcal{X},\mathcal{Y}}$ has the expected dimension for many choices of \mathcal{X} and \mathcal{Y} . On the other hand, as Hadamard product of naïve Bayes models, $\text{RBM}_{\mathcal{X},\mathcal{Y}}$ has dimension less than expected whenever for some $j \in [m]$ the $|\mathcal{Y}_j|$ -mixture of $\mathcal{E}_{\mathcal{X}}$ has dimension less than expected (Proposition 19).

Various questions remain unsettled: What is the exact dimension of the naïve Bayes models with general discrete variables? What is exactly the smallest number of hidden variables that make an RBM a universal approximator? Does a binary RBM always have the expected dimension? The geometric-combinatorial picture presented in this paper may be helpful in solving these problems.

Acknowledgments

The first author is grateful to Nihat Ay and Johannes Rauh. Part of this work was accomplished while he visited the Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany, in September and October 2012. This work is supported in part by DARPA grant FA8650-11-1-7145.

References

- [1] M. Aoyagi. Stochastic complexity and generalization error of a Restricted Boltzmann Machine in Bayesian estimation. *J. Mach. Learn. Res.*, 99:1243–1272, August 2010.
- [2] Y. Bengio. Learning deep architectures for AI. *Found. Trends Mach. Learn.*, 2(1):1–127, 2009.
- [3] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 153–160. MIT Press, Cambridge, MA, 2007.
- [4] M. A. Carreira-Perpiñán and G. E. Hinton. On contrastive divergence learning. In *Proceedings of the 10-th International Workshop on Artificial Intelligence and Statistics*, 2005.
- [5] M. V. Catalisano, A. V. Geramita, and A. Gimigliano. Secant varieties of $\mathbb{P}^1 \times \cdots \times \mathbb{P}^1$ (n -times) are not defective for $n \geq 5$. *J. Algebraic Geometry*, 20:295–327, 2011.
- [6] G. Cohen, I. Honkala, S. Litsyn, and A. Lobstein. *Covering Codes*. North-Holland Mathematical Library. Elsevier Science, 2005.
- [7] M. A. Cueto, J. Morton, and B. Sturmfels. Geometry of the restricted Boltzmann machine. In M. A. G. Viana and H. P. Wynn, editors, *Algebraic methods in statistics and probability II, AMS Special Session*, volume 2. American Mathematical Society, 2010.
- [8] G. E. Dahl, R. P. Adams, and H. Larochelle. Training restricted Boltzmann machines on word observations. *arXiv:1202.5695*, 2012.
- [9] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.
- [10] Y. Freund and D. Haussler. Unsupervised learning of distributions of binary vectors using 2-layer networks. In J. E. Moody, S. J. Hanson, and R. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 912–919. Morgan Kaufmann, 1991.
- [11] E. Gilbert. A comparison of signalling alphabets. *Bell System Technical Journal*, 31:504–522, 1952.
- [12] G. E. Hinton. Products of experts. In *Proceedings 9-th ICANN*, volume 1, pages 1–6, 1999.
- [13] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.
- [14] G. E. Hinton. A practical guide to training restricted Boltzmann machines, version 1. Technical report, UTM12010-003, University of Toronto, 2010.
- [15] G. E. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.
- [16] N. Le Roux and Y. Bengio. Representational power of restricted Boltzmann machines and deep belief networks. *Neural Computation*, 20(6):1631–1649, 2008.
- [17] P. M. Long and R. A. Servedio. Restricted Boltzmann machines are hard to approximately evaluate or simulate. In J. Fürnkranz and T. Joachims, editors, *ICML*, pages 703–710. Omnipress, 2010.
- [18] D. Lowd and P. Domingos. Naive Bayes models for probability estimation. In *Proceedings of the Twenty-second International Conference on Machine Learning*, pages 529–536. ACM Press, 2005.

- [19] T. K. Marks and J. R. Movellan. Diffusion networks, products of experts, and factor analysis. In *Proc. 3rd Int. Conf. Independent Component Anal. Signal Separation*, pages 481–485, 2001.
- [20] R. Memisevic and G. E. Hinton. Learning to represent spatial transformations with factored higher-order Boltzmann machines. *Neural Computation*, 22(6):1473–1492, June 2010.
- [21] G. Montúfar. Mixture decompositions of exponential families using a decomposition of their sample spaces. *Kybernetika*, 49(1), 2013.
- [22] G. Montúfar and N. Ay. Refinements of universal approximation results for deep belief networks and restricted Boltzmann machines. *Neural Computation*, 23(5):1306–1319, 2011.
- [23] G. Montúfar and J. Morton. When does a mixture of products contain a product of mixtures? *Preprint available at <http://arxiv.org/abs/1206.0387>*, 2012.
- [24] G. Montúfar and J. Rauh. Scaling of model approximation errors and expected entropy distances. In *Proc. of the 9th Workshop on Uncertainty Processing (WUPES 2012)*, pages 137–148, 2012. Preprint available at <http://arxiv.org/abs/1207.3399>.
- [25] G. Montúfar, J. Rauh, and N. Ay. Expressive power and approximation errors of restricted Boltzmann machines. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 415–423, 2011.
- [26] S. Osindero and G. E. Hinton. Modeling image patches with a directed hierarchy of Markov random fields. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1121–1128. MIT Press, Cambridge, MA, 2008.
- [27] L. Pachter and B. Sturmfels. Tropical geometry of statistical models. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46):16132–16137, Nov. 2004.
- [28] M. Ranzato, A. Krizhevsky, and G. E. Hinton. Factored 3-way restricted boltzmann machines for modeling natural images. *Journal of Machine Learning Research - Proceedings Track*, 9:621–628, 2010.
- [29] W. E. Roth. On direct product matrices. *Bulletin of the American Mathematical Society*, 40:461–468, 1934.
- [30] R. Salakhutdinov, A. Mnih, and G. E. Hinton. Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning, ICML '07*, pages 791–798, New York, NY, USA, 2007. ACM.
- [31] T. J. Sejnowski. Higher-order Boltzmann machines. In *Neural Networks for Computing*, pages 398–403. American Institute of Physics, 1986.
- [32] P. Smolensky. Information processing in dynamical systems: foundations of harmony theory. In *Symposium on Parallel and Distributed Processing*, 1986.
- [33] T. Tran, D. Q. Phung, and S. Venkatesh. Mixed-variate restricted Boltzmann machines. In *Proc. of 3rd Asian Conference on Machine Learning (ACML)*, pages 213–229, 2011.
- [34] R. Varshamov. Estimate of the number of signals in error correcting codes. *Doklady Akad. Nauk SSSR*, 117:739–741, 1957.
- [35] M. Welling, M. Rosen-Zvi, and G. E. Hinton. Exponential family harmoniums with an application to information retrieval. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1481–1488. MIT Press, Cambridge, MA, 2005.